Korelace

Kde jsme

- Souvislost kategorických proměnných → crosstab
- Souvislost kategorické a kardinální proměnné → srovnání průměrů
- Souvislost kardinálních proměnných → korelační koeficient

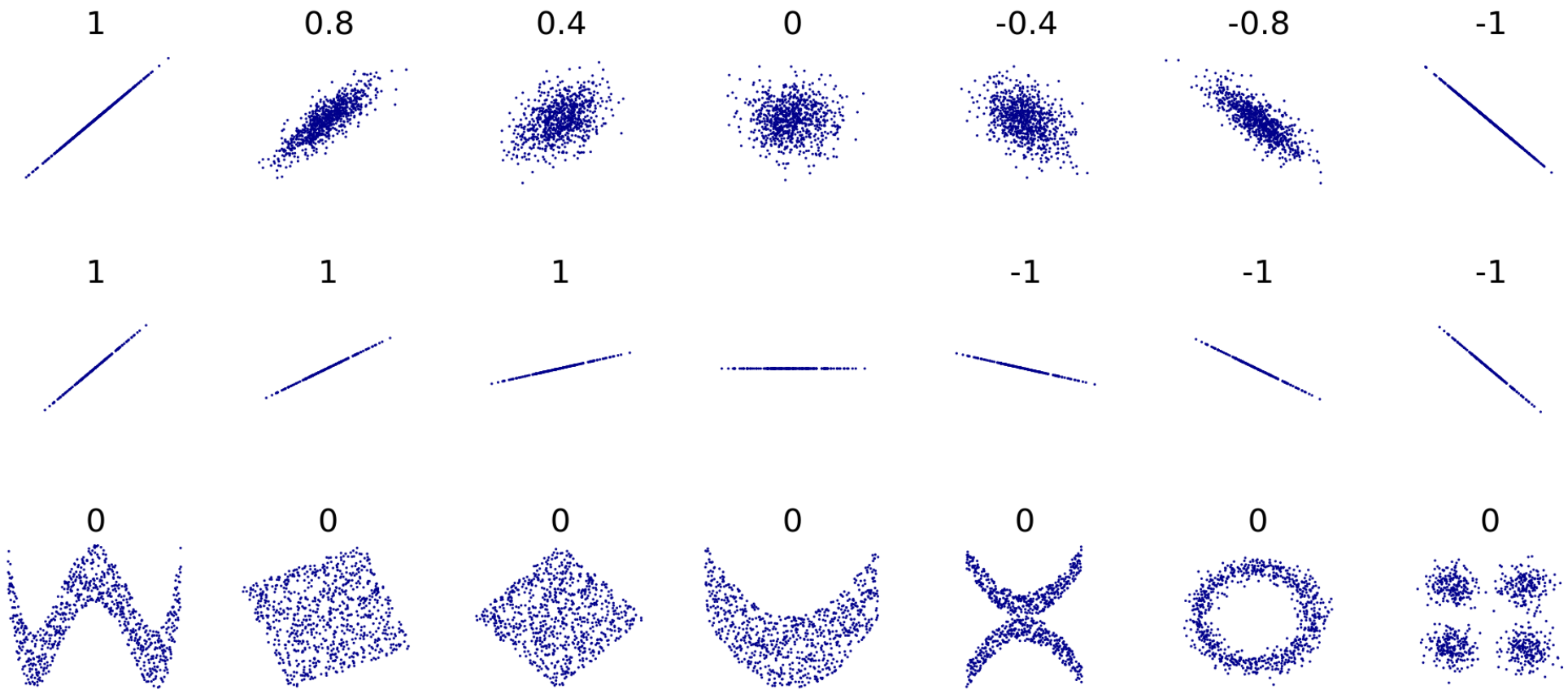
Korelace

- Jak se pozná souvislost:
  - S růstem hodnot jedné proměnné narůstají nebo klesají hodnoty druhé proměnné
- Lineární vztah: nárůst/pokles musí být pro nízké i vysoké hodnoty proměnné stejný
- Typy korelačních koeficientů:
  - Pearson
  - Spearman
  - Kendall

Korelace

- Hodnoty koeficientu:
  - Rozsah od -1 po 1
  - +1 = perfektní kladná souvislost – s růstem jedné proměnné roste druhá
  - -1 = perfektní záporná souvislost – s růstem jedné proměnné klesá druhá
  - 0 = žádná souvislost

- Čím více je hodnota vzdálena od nuly, tím je souvislost silnější
- Existují „tabulky“ k hodnocení síly vztahu



Autor: DenisBoigelot, original uploader was Imagecreator, CC0, https://commons.wikimedia.org/w/index.php?curid=15165296

Korelace

- Hodnoty koeficientu:
  - Rozsah od -1 po 1
  - +1 = perfektní kladná souvislost – s růstem jedné proměnné roste druhá
  - -1 = perfektní záporná souvislost – s růstem jedné proměnné klesá druhá
  - 0 = žádná souvislost

- Čím více je hodnota vzdálena od nuly, tím je souvislost silnější
- Existují „tabulky“ k hodnocení síly vztahu

- Je vhodné si data vizualizovat

Pearsonův korelační koeficient

- Parametrická operace
- Předpoklady:
  - Kardinální data (výjimka - jedna z proměnných může být dichotomická)
  - Normální rozložení / dostatečná velikost vzorku (min. 200-500)
- Značení - R
- Citlivost na odlehlé případy

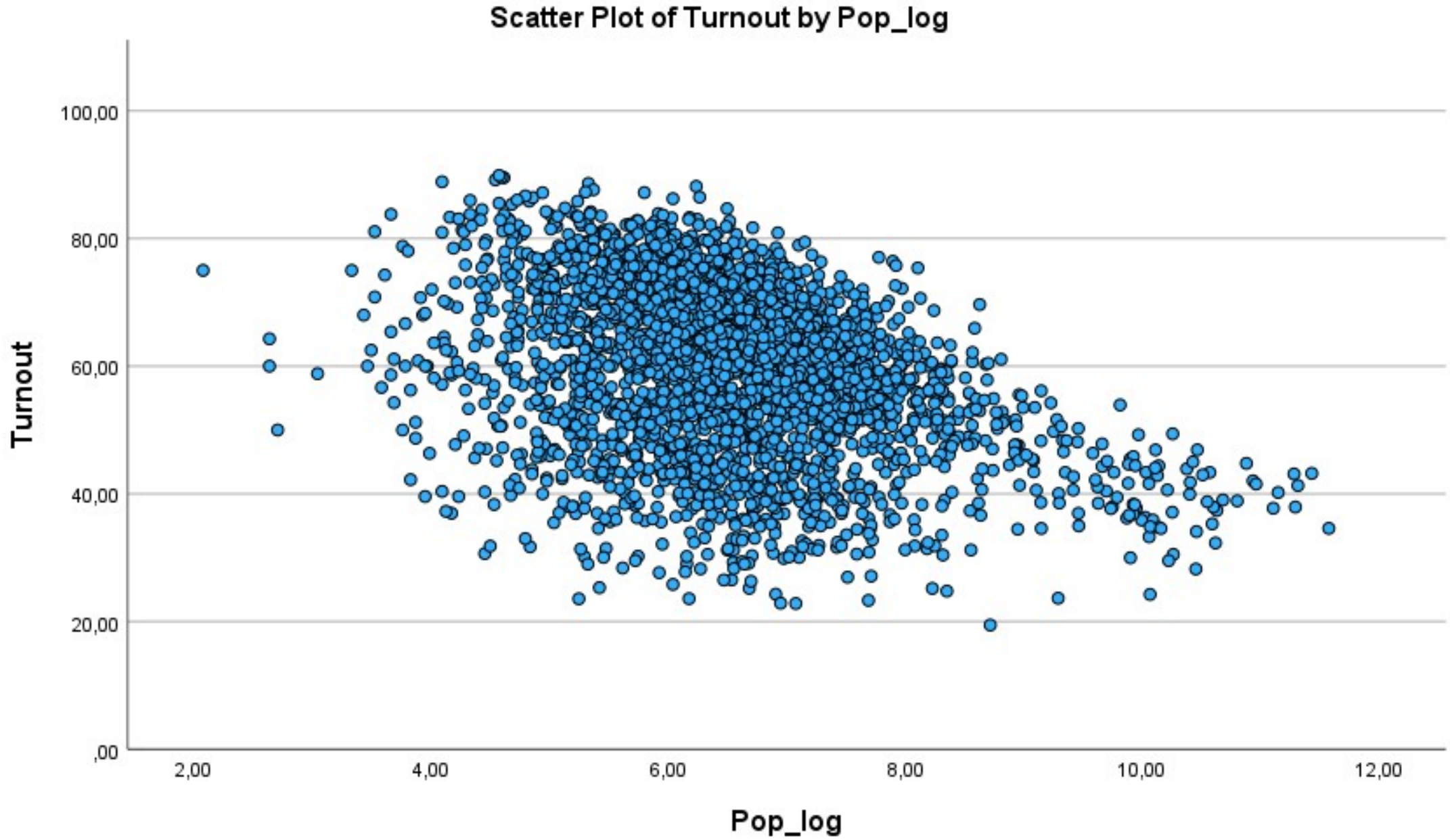
Práce v SPSS

- Analyze → Correlate → Bivariate:
  - Zvolit proměnné
  - Pearsonův koeficient je přednastavený
  - Pro sledování signifikance zvolit *Flag significant correlations*

- Options:
  - Možnost spočítat základní statistiky
  - Vynechání hodnot / případů
    - Listwise – pokud počítáme více korelačních koeficientů, všechny budou založeny na stejných datech
    - Pairwise – missing odstraněny zvlášť v každém páru

Příklad

- Existuje souvislost mezi počtem obyvatel obcí a volební účastí?
- Máme data **ze všech** měst v zemi
- Co potřebujeme zjistit:
  - Je mezi proměnnými vztah?
  - Je statisticky významný? Je to vůbec důležité?
  - Má výsledek věcný význam?


Scatter Plot of Turnout by Pop_log
Turnout
100,00
80,00
60,00
40,00
20,00
,00
2,00
4,00
6,00
8,00
10,00
12,00
Pop_log

Correlations

		Pop_log	Turnout
Pop_log	Pearson Correlation	1	-,366**
	Sig. (2-tailed)		,000
	N	2926	2919
Turnout	Pearson Correlation	-,366**	1
	Sig. (2-tailed)	,000	
	N	2919	2919

**. Correlation is significant at the 0.01 level (2-tailed).

Pearsonův korelační koeficient

- Síla vztahu:
  - ± 0,1 – slabý
  - ± 0,3 – střední
  - ± 0,5 – silný

- Spíše informativní hodnoty (mezi r = 0,49 a r = 0,51 žádný zásadní rozdíl není)

Pearsonův korelační koeficient

- Výjimka z kardinálních dat → korelace jedné kardinální proměnné a jedné dichotomické
- Tzv. point-biserial korelace
- Kladné / záporné výsledné hodnoty závisí na kódování dichotomické proměnné

Populace obcí (2 kategorie) a účast

Correlations

		Turnout	Pop_2cat
Turnout	Pearson Correlation	1	-,294**
	Sig. (2-tailed)		<,001
	N	2919	2919
Pop_2cat	Pearson Correlation	-,294**	1
	Sig. (2-tailed)	<,001	
	N	2919	2926

**. Correlation is significant at the 0.01 level (2-tailed).

Populace obcí (2 kategorie) a účast

Correlations

		Turnout	Pop_2cat_inv
Turnout	Pearson Correlation	1	,294**
	Sig. (2-tailed)		<,001
	N	2919	2919
Pop_2cat_inv	Pearson Correlation	,294**	1
	Sig. (2-tailed)	<,001	
	N	2919	2926

**. Correlation is significant at the 0.01 level (2-tailed).

Kardinální a dichotomická proměnná

- Můžeme použít Pearsonův korelační koeficient
- Máme (vhodnější) alternativu?

Neparametrická korelace

- Využívá pořadí případů, nikoli samotné hodnoty proměnné

- **Spearmanovo Rho:**
  - Využíván zejména pro kombinaci ordinálních proměnných
  - V menších vzorcích (do 200-500) při porušení normality
  - Výsledné hodnoty jsou ve stejném pásmu jako u PKK (od -1 po 1)

- **Kendallovo Tau:**
  - Pro malé vzorky a menší množství kategorií
  - Volba mezi kendallem a crosstabem
  - Některé hodnoty se velice často opakují

SPSS

- Analyze → Correlate → Bivariate:
  - Zvolit proměnné
  - Vybrat *Spearman a/nebo Kendall*

Correlations

			How happy are you	How often socially meet with friends, relatives or colleagues
Kendall's tau_b	How often socially meet with friends, relatives or colleagues	Correlation Coefficient	,153**	
		Sig. (2-tailed)	<,001	
		N	2351	
	How many people with whom you can discuss intimate and personal matters	Correlation Coefficient	,075**	,244**
		Sig. (2-tailed)	<,001	<,001
		N	2345	2347
Spearman's rho	How often socially meet with friends, relatives or colleagues	Correlation Coefficient	,191**	
		Sig. (2-tailed)	<,001	
		N	2351	
	How many people with whom you can discuss intimate and personal matters	Correlation Coefficient	,093**	,298**
		Sig. (2-tailed)	<,001	<,001
		N	2345	2347

**. Correlation is significant at the 0.01 level (2-tailed).

Interpretace výsledků

- Základní pravidlo – **korelace ≠ kauzalita**
- Korelace vyjadřuje pouze souvislost mezi proměnnými, neukazuje na žádnou příčinu a následek
- Vliv třetích proměnných
- Nemožnost konstatovat kauzalitu i pokud se jeví jako logická
- Statistické zjištění nemá automaticky věcný význam

Number of Grand Slam Finals played by Roger Federer

correlates with

The number of electronics engineers in New Mexico

Number of Grand Slam Finals played by Roger Federer · Source: Wikipedia

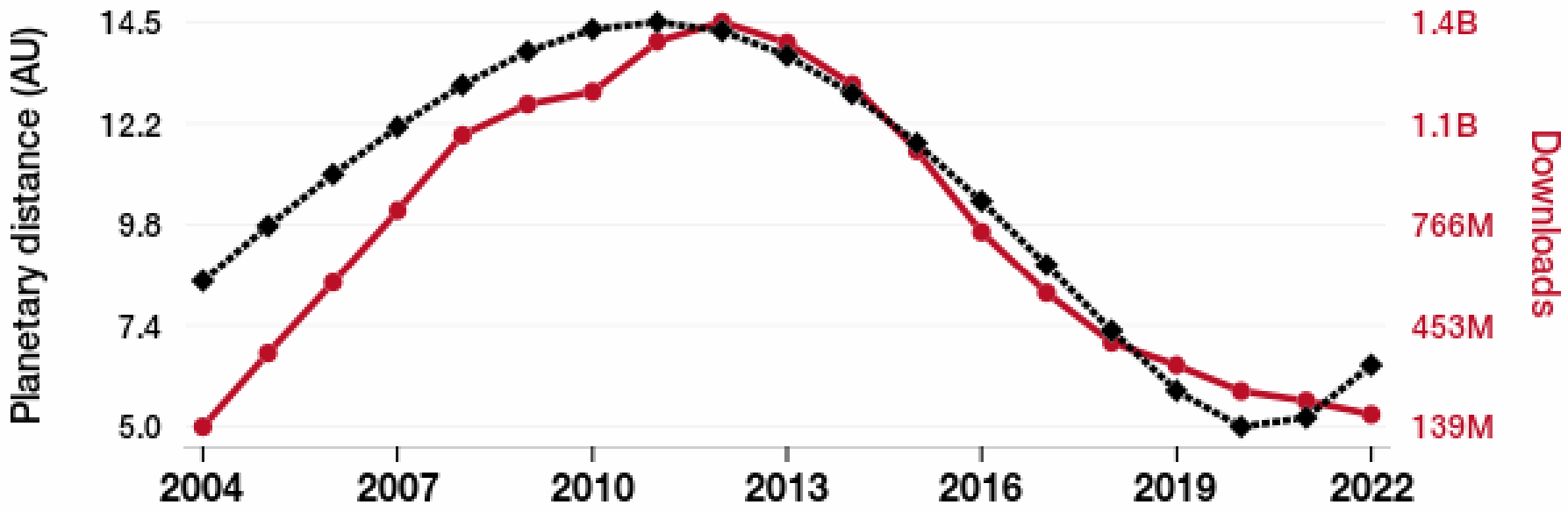
BLS estimate of electronics engineers, except computer in New Mexico · Source: Bureau of Larbor Statistics

2003-2015, $r=0.905$, $r^2=0.819$, $p<0.01$ · tylervigen.com/spurious/correlation/1077

The distance between Saturn and Jupiter

correlates with

Total Digital Music Single Downloads



The average distance between Saturn and Jupiter as measured on the first day of each month · Source: Caclculated using Astropy

Total Digital Music Single Downloads · Source: Statista

2004-2022, r=0.932, r^2=0.869, p<0.01 · tylervigen.com/spurious/correlation/1163

Práce s koeficienty

- Co uvádět:
  - Jaký koeficient byl použit, kolik případů bylo v analýze
  - V tabulce: hodnoty koeficientu, hvězdičky a sig. jen pokud je potřeba zohlednit signifikanci
  - K hvězdičkám je nutné dodat legendu